Fast Detection of Transformed Data Leaks Framework on Mail Server

KavithaPriya.C. J[1], Deepika.M[2], Shalini.k[3], Suganya.S[4]

kavithapriyacj@jeppiaarinstitute.org , deepikamalairaju@gmail.com , sugan07srini@gmail.com , k.shalini24696@gmail.com

Assistant Professor, Dept. of IT, Jeppiaar Institute of Technology[1]

B. Tech Student, Dept. of IT, Jeppiaar Institute of Technology[2,3,4]

ABSTRACT: Organizations in particular faces a major threat of data leakage of sensitive information. Data leakages most commonly occur due to human error, deliberately planned attacks and forwarding of Emails to unauthorized or third party. Thus the detection and prevention of data leaks is a prime focus in the area of security. However, detecting the exposure of sensitive information is challenging due to data transformation in the content. To do this we would require various solutions that detect the malware planted deliberately to seek out information and checks for the data sent through mail. These data leaks are handled by the DLD [data leak detection] server that is equipped within every organization. In our paper, we use the Apache Lucene Framework – a java based indexing and searching technology, which is used for spellchecking, fuzzy querying, pattern matching etc. In addition to this we Levenshtein distance to check two different sequence of strings and a shuffling algorithm that is enabled to match two shuffled strings of sensitive information.

KEYWORDS: Data leaks, Data leak detection, Shuffling, Lucene, Pattern matching, Levenshtine distance.

I.INTRODUCTION

Data leaks have been a serious threat towards the security of an organization. It may be caused due to human errors or intended data leaks due to malicious software's. Recent reports on data leaks shows the number has gone up 10 times in the last 4 years. The prevention of these data leaks is the primary concern of any organization. In order to prevent data leaks the organizations have been using the concept of direct filtration which filters the mail with the confidential and sensitive contents.

II. LITERATURE SURVEY

a. DLD and DLP server

In [1] A Literature Review on Cyber Forensic and its Analysis tools Mandeep Kaur, Navreet Kaur, Suman Khurana (2016), paved way to becoming an insight on what is Cyber forensics and how it can be used to prevent data leaks. Due to the frequent use of internet technologies cyber-attacks can occur. Digital forensic is opted for acquiring electronic information and investigation of malicious evidence found in system or on network in such a manner that makes it admissible in court. It is also used to recover lost information in a system. The recovered information is used to prosecute a criminal. Number of crimes committed against an internet and malware attacks over the digital devices have increased. Memory analysis has become a critical capability in digital forensics because it provides insight into the system state that should not be represented by traditional media analysis.

There are various forensic analysis tools present that are opted for different purposed. They are: Autopsy is a digital forensics platform, operating system forensics is a toolkit that provides lots of information, RAM forensics is forensic analysis of a computer's memory dump, Digital

Forensics Framework is an Open Source computer forensics software, Wire shark is free open-source packet analyzer and True Crypt is an open source-available freeware utility used for encryption.

The various accountability of data leaks has been anticipated by Data Breach Quick View: 2014 Data Breach Trends. The accounts of fraud, breaches and hacking have been increased rapidly over the recent years [3]. Moreover, the amount incidents that were exposed varied among different sectors like business, government, medical and Education. The amounts of security breach in the business sector increased quantitatively around the years. The analysis of the incidents around the year 2014, exposed various breach types like hacking, fraud SE, web etc. Among these, hacking was the most commonly used breach type along with outside threats taking the first place. There were 87 countries reporting at least one data breach in 2014.

In [6] Email Security: Defending the Enterprise, Trend Micro, Kaspersky lab and Symantec collaborate their ideas together in the process of preventing data leaks in the organization using various concepts. It is a proven fact that in this world that runs along the line of technology, is using mails as a daily part of their communication. In addition, it is also a fact that a recent study conducted by McKinsey revealed that workers spent 28% of their time in reading and reply to mails from their working partners. There are a lots of threats that acts as a stimulus to leaks of data. Some of them are viruses, worms, Trojan horses, malicious codes, Insider threats and Inappropriate content.

In order to prevent the leaks of data organizations employs an outsourcing server that handles the data leaks of the mail server using techniques like content filtering, encryption and data loss prevention. There are various solutions that are integrated by certain vendors like McAfee, Symantec and Trend Micro.

In addition to this, in [7] Should You Outsource Email, Brien M. Posey talks about the Microsoft Exchange server that is used to filter both the incoming and the outsourced mails of an organization. It is directly connected to the mail server and performs a five staging process that is needed to check the contents of the mail and identify them as data leaks or stimulus for a data leak.

Apart from filtering the outsourced mails, we have filters that identify the threats from outside, that are prevented using filters and Anti spammers. In [8] Cisco Outbreak Filters (2012), we discuss about how to tackle attackers smartly and by staying ahead of them using sharpened tactics. These outbreak filters are built upon the virus filters, that not only scans the content of the email but also scans the URL and process them in real time.

The steps involved are the incoming mail is scanned by the outbreak server, then the URL is rewritten in the email. Next this email is forwarded to the intended person and redirects it to the public proxy that scans the mail and informs about any malicious code. In case of any malware or malicious code, the page is blocked by the server and a block page message is sent to the server. If the mail is of such importance, then the user is provided with choices of using the URL in the proxy or going to the website directly.

b. Algorithms

Now in the recent trends, the organizations are using certain algorithms that enable the prevention of data leaks. In [2] Rapid screening of transformed data leaks with efficient algorithms and parallel computing, X. Shu, J. Zhang, D. Yao, and W.-C. Feng (2015), gave the basic idea of exposing the transformed data leaks. In the previous systems, a method called set intersection was used to identify the set of content files. But this was an order less method. Hence an oblivious sampling alignment and comparison algorithm was introduced to overcome this. Taking this into consideration, the next system proposed had additional features to detect the amount of sensitive information possessed. Now, we are designing a system where these data leaks can be detected and prevented on mail servers and web services.

Prevention of inadvertent data leaks enable that the contents in the file system or the network traffic is available for the inspection using channels. Preventing information leakage from indexing in the cloud, A. Squicciarini, S. Sundareswaran, and D. Lin

(2010), they talk about how data can be prevented over the cloud systems [5]. Cloud computing enables highly scalable services those that can be easily consumed and obtained over the Internet on an as-needed basis.

The most common scheme for supporting efficient search over distributed content is to build a centralized inverted index. The index maps each term to the document that contains the search term, and is queried by the user to obtain a list of matching documents. This scheme is usually adopted by web search engines and mediators. They introduce a distributed access-control enforcing protocol, which differs from our approach in the assumptions and the type of architecture employed.

In [4] Data protection models for service provisioning in the cloud, D. Lin and A. Squicciarini, the concept of the cloud includes a number of implementation based on the service they provide from application service provisioning grid and utility computing to software as a service. The organization or enterprises processes remotely unknown matching users that do not own or operate. This approach comes with privacy and security risks. The barriers to the adoption of cloud service is that the users fear of confidential data leakage and loss of privacy in the cloud.

This approach is to protect a user data over the cloud. During the service provisioning the most challenging problem is to ensure the user's data. To address these challenges the data protection in the cloud approaches uses policy techniques and framework that are used in data protection. In the cloud they are: Policy ranking that helps users quickly identify a suitable policy provider, Automatic policy generation that takes policies and requirement from the users, and Policy enforcement that enforces a privacy policy across multiple parties.

Amazon,Google,Microsoft,Saleforce.com and Sun are cloud providers. They are the small portion of service provider in the cloud. Moving on to data protection models the participating party may need to update the privacy policies; a service provider may need to transfer its operation together with the user's data.

The policy ranking models aims to find the service provider with the similar privacy policies compared to the user privacy requirement. The two important factors that are required are Privacy and Efficiency requirement. Based on these two factors three policy ranking models are established. They are User oriented ranking model, Service provider oriented ranking model, and Worker based ranking model. These are all the factors and policies are used to protect the data's in the cloud.

III. RELATED WORKS

In [9] We focus on a new approach towards identity detection, that uses a twostep strategy of first finding out the objective of the user and whether the aim matches along with the user interactions. In order to notify and verify this, we use a framework that captures the user intent and monitors along with the user interactions. This framework also exhibits a policy of "WYSIWYS" (What You See Is What You Send) to confirm the text captured from the user's intent is same as what is present in the text based application.

This framework is controlled and monitored for traffic analysis, and thus performs and entirely different route of steps once traffic is notified to the gyrus frame work. This has been used for various applications like Digsby: Yahoo Messenger and Twitter, to create a user intent signature, etc. Now in our project we tend to use a framework called as Lucene framework, that also helps in the capturing of text, monitoring the data and indexing, patterning them into various different aspects that helps in the prevention of data leakage attacks.

In [10] we discuss about the various methods that can be used to measure and quantify outbound data leaks. Outbound data leaks may mostly occur when data is being outsourced. These data leaks have been identified by the systems that helps us to track the data leaks in the server. But, the major disadvantage of this being that we cannot track the leak of data that is encrypted. To overcome this, they introduced, certain algorithms that was successfully accepted into making it useful to identify the outbound data leaks. But along with this, the most promising challenges of outbound leaks is find them in the Email servers.

Thus, in our paper we tend to identify the data leaks that can happen within an organiasation, with help of Email server. This is being protected using the Lucene framework and the levinshtine distance algorithm, along with a shuffling algorithm.

IV PROPSED SYSTEM

In our proposed system, we enable a Lucene framework that is integrated with the mail server, which acts as a proxy, in filtering the outsourced mails. This framework is a java based framework that aids in searching, indexing, pattern matching, reasoning etc. This framework enables the proxy to index the sensitive information, and then every time a mail is outsourced to another domain, then the contents are searches, patterns are matched with the indices already made by the framework. Now, using two algorithms, levenshtine algorithm and shuffling algorithm, we and match the strings of the content and sensitive information even if they are not in order.

A. Lucene Framework

In [11] it briefly explains about how the full text is searched and how it is retrieved. For this Lucene framework is used. In the rapid growth and development of the internet there is vast amount of information that need to be searched. The full text search includes the search of text, images, videos etc. The process of the full text search starts from the building a text database in this it stores all the data that are retrieved by the user. Then creating the index in this the indexing will improve the speed of information. After the indexing searching is started, user can search any information that is needed. Last is the filter and sorting in this certain rules are followed and then the information is given to the user. Lucene structure has a strong object-oriented feature and written in Java. It has 5 main modules like analysis, document, index, search and query parser. In a Lucene the index is made up of segments and the segments is made up of documents and is documents is made up of fields. Initially a infect is starts from the add Document method of Index Writer.

In [12] a comparison of open source search engine development frameworks in the context of their malleability for constructing multilingual search index. There are lots of open source frameworks available to build a search engine. These open source frameworks provide multiple features to build an inverted index of web documents used for information retrieval.

Some features like Scalability, term storage, document posting list storage etc., are common across these frameworks.

These frameworks facilitate customization of building index to make it compatible for the desired application. To retain the structure of a document in an inverted index, field based indexing is used. Instead of viewing a document as a collection of terms, the document is viewed as a collection of fields and the field as a collection of terms. Each document that needs to be indexed is parsed and terms in the document are grouped into fields prior to indexing. There are lots of works on building the inverted index using an open source framework.

The most popular indexing library is Apache Lucene (Apache Lucene, 2011). Lucene is not a complete search engine framework, but an indexing library used to generate inverted index from crawled documents. Lucene needs to be plugged in with a crawler in order to index web documents. Apache Lucene provides facilities for customizing the library and makes it easily pluggable with the crawler that is being used.

B. Levenshtine Distance

Levenshtine distance is named after the Russian scientist Vladimir Levenshtein, who devised the algorithm in 1965.

Levenshtein distance (LD) is a measure of the similarity between two strings, one is the source string (s) and the target string (t). The distance calculate through the algorithm is the number of deletions, insertions, or substitutions required to transform s into t. For example,

- If s is "test" and t is "test", then LD(s,t) = 0, because no transformations are needed, thus the strings are identical.
- If s is "test" and t is "tent", then LD(s,t) = 1, because one substitution is sufficient to transform s into t, thus the strings are non-identical.

The greater the Levenshtein distance, the more different the strings are. The Levenshtein distance algorithm has been used in various applications like:

1. Spell checking
2. Speech recognition
3. DNA analysis
4. Plagiarism detection

ALGORITHM: LEVENSHTINE DISANCE

STEPS:

```
1. a) Set n to be the length of s.
   b) Set m to be the length of t.
   c) If n = 0, return m and exit.
   d) If m = 0, return n and exit.
Construct a matrix containing 0...m rows and 0...n columns.
```

2. a) Initialize the first row to 0...n.
   b) Initialize the first column to 0...m.
3. Examine each character of s (i from 1 to n).

4. Examine each character of t (j from 1 to m).

5. a) If s[i] equals t[j], the cost is 0.
   b) If s[i] doesn't equal t[j], the cost is 1.

6. Set cell d[i,j] of the matrix equal to the minimum of:
a. The cell immediately above plus 1: d [i-1, j] + 1.
b. The cell immediately to the left plus 1: d [i, j-1] + 1.
c. The cell diagonally above and to the left plus the cost: d [i-1, j-1] + cost.
7. After the iteration steps (3, 4, 5, 6) are complete, the distance is found in cell d [n, m].

C. Shuffling Algorithm

This algorithm finds the piece of content that have been shuffled and are matched and indexed with the source sensitive content. This is used to reduce the date leaks that can be enabled due to disordering of the sequential alignment of the source content. Due to the usage of shuffling algorithm, levenshtine distance and the Lucene framework, we are able to reduce the traffic flow intensity when compared to that of DLD servers.

D. Traffic intensity analysis

When comparing the two the normal procedure of filtering and the filtering using Lucene framework, we can come to conclude that the rate of traffic in the network when DLD is used is comparatively higher than that of traffic flow analyzed in Lucene framework.

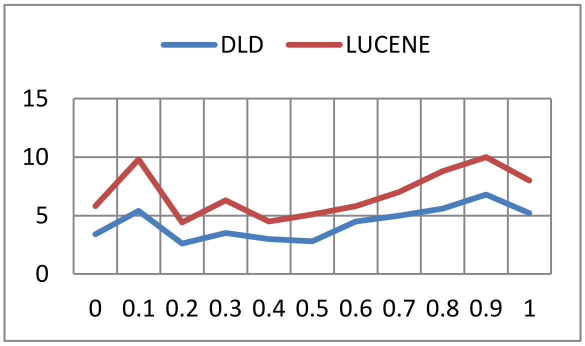

Fig: Comparison of traffic intensity between the DLD and the Lucene framework.

E. Sensitivity Threshold

There are two methods that are used to measure the sensitivity of the data and analyze the sensitivity data leaks. The AlignDLD method provides a maximum threshold of 0.2 and the Coll-Inter method overlaps the sensitive information and thus no data leaks are observed in it. It has a low accuracy power than that of AlignDFD but has high false positive rates.

The random and pervasive substitution method that is used produces a sensitivity threshold of 0.3 along with a 80% prevention of leaks.

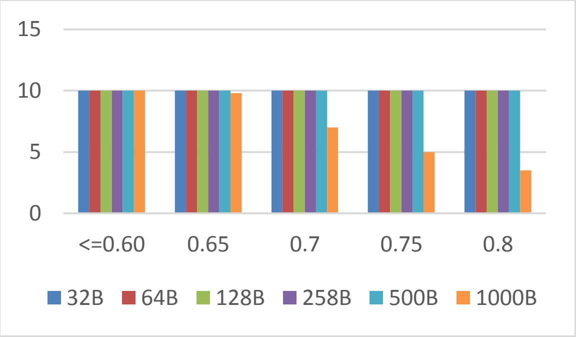

Fig: Shows the partial data leaks in AlignDLD method that is used to measure the sensitivity threshold of the data outsourced along with the sensitivity threshold.

V CONCLUSION

In this paper, we studied the data leakage prevention for mail servers, and found out that using frameworks data leaks can be prevented in the same. We, used an open source framework called Apache Lucene framework, which is based on the java platform and enables the pattern matching, searching, and indexing. In addition, we also use levenshtine distance and shuffling algorithm that enhances the concept of pattern matching in unordered strings. Moreover, we also estimated the traffic intensity during prevention of data leaks in traditional DLP and using Lucene framework. In our paper we have studied for data leaks in text documents. Further in future, we can study and analyze the prevention of data leaks using images or steganography.

VI REFERENCES

[1] Mandeep Kaur1, Navreet Kaur2, Suman Khurana3, “A Literature Review on Cyber Forensic and its Analysis tools”, in *International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016*

[2] X. Shu, J. Zhang, D. Yao, and W.-C. Feng, “Rapid screening of transformed data leaks with efficient algorithms and parallel

computing," in *Proc. 5th ACM Conf. Data Appl. Secur. Privacy (CODASPY)*, San Antonio, TX, USA, Mar. 2015, pp. 147–149.

[3] (Feb. 2015). *Data Breach QuickView: 2014 Data Breach Trends*. [Online]. Available: https://www.riskbasedsecurity.com/reports/2014-YEDataBreachQuickView.pdf, accessed Feb. 2015.

[4] D. Lin and A. Squicciarini, "Data protection models for service provisioning in the cloud," in *Proc. 15th ACM Symp. Access Control Models Technol.*, 2010, pp. 183–192.

[5] A. Squicciarini, S. Sundareswaran, and D. Lin, "Preventing information leakage from indexing in the cloud," in *Proc. 3rd IEEE Int. Conf. Cloud Comput.*, Jul. 2010, pp. 188–195.

[6] Email Security: Defending the Enterprise, Trend Micro, Kaspersky lab and Symantec – White Paper.

[7] Should You Outsource Email, Brien M. Posey – White Paper.

[8] Cisco Outbreak Filters (2012) – White Paper by CISCO.

[9] Y. Jang, S. P. Chung, B. D. Payne, and W. Lee, "Gyrus: A framework for user-intent monitoring of text-based networked applications," in *Proc.23rd USENIX Secur. Symp.*, 2014, pp. 79–93.

[10] K. Borders and A. Prakash, "Quantifying information leaks in outbound Web traffic," in *Proc. 30th IEEE Symp. Secur. Privacy (SP)*, May 2009, pp. 129–140.

[11] Rujia Gao, Danying Li, Wanlong Li, Yaze Dong, "Application of Full Text Search Engine Based on Lucene" in ***Advances in Internet of Things*, 2012, 2, 106-109** http://dx.doi.org/10.4236/ait.2012.24013 Published Online October 2012 (http://www.SciRP.org/journal/ait)

[12] Arjun Atreya V, Swapnil Chaudhari1 Pushpak Bhattacharyya, Ganesh Ramakrishnan, "Building Multilingual Search Index using open source framework"

[13] http://www.levenshtein.net/index.html

[14]http://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Fall2006/Assignments/editdistance/Levenshtein%20Distance.htm